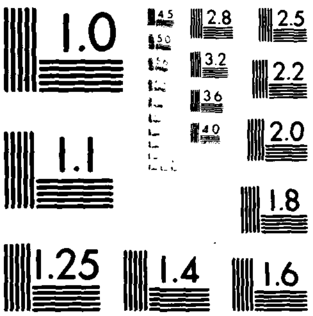
AD-A092 610 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
LONG COMMON SUBSEQUENCES AND THE PROXIMITY OF TWO RANDOM STRING--ETC(U)
NOV 80 J M STEELE N00014-76-C-0475
UNCLASSIFIED TR-293 NL

1 OF 1
AD A
092610

END
DATE
FILMED
1-81
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF

6

LONG COMMON SUBSEQUENCES AND THE PROXIMITY OF TWO RANDOM STRINGS

By

10 J. MICHAEL STEELE

14 TR-

9 TECHNICAL REPORT NO. 293

11 6 Nov 80

NOVEMBER 6, 1980

12 19

15 Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATSITICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

332580

LONG COMMON SUBSEQUENCES AND THE PROXIMITY OF TWO RANDOM STRINGS

By

J. Michael Steele

I. INTRODUCTION

Long molecules such as proteins and nucleic acids can be thought of schematically as sequences from a finite alphabet $\mathcal{Q}$. From an evolutionary point of view it is natural to compare molecules by considering their common ancestors, and in schematic terms this reduces to the problem of considering the longest common subsequence of two given sequences.

Sankoff (1972) gave an efficient algorithm for calculating the length of the longest common subsequence. Subsequently, Sankoff and Cedergren (1973), and Sankoff, Cedergren, and Lapalme (1976) considered a number of empirical cases and conducted some Monte Carlo investigations. The first formal probabilistic analysis of the problem of long common subsequences was initiated in Chvátal and Sankoff (1975). To describe their work we first introduce some notation.

By X_i and X'_i, $1 \leq i < \infty$, we denote two sequences of independent, and identically distributed random variables with values in $\mathcal{Q}$. The random variable of main interest is

$$L_n := \max\{k: X_{i_1} = X'_{j_1}, X_{i_2} = X'_{j_2}, \ldots, X_{i_k} = X'_{j_k} \text{ where}$$

$$1 \leq i_1 < i_2 < \cdots < i_k \leq n \text{ and } 1 \leq j_1 < j_2 < \cdots < j_k \leq n\} .$$

In words, L_n is the largest cardinality of any subsequence common to the sequences $(X_1,X_2,\ldots,X_n)$ and $(X_1',X_2',\ldots,X_n')$.

Under the assumption that $|G| = k$ and that X_i and X_i' are both uniform on G, Chvátal and Sankoff proved the existence of the limit of the means,

$$\lim_{n \to \infty} EL_n/n = c_k . \tag{1.1}$$

Among other results, Chvâtal and Sankoff obtained upper and lower bounds on c_k. These authors proved no results for Var L_n, but on the basis of a Monte-Carlo study they were lead to conjecture Var $L_n = o(n^{2/3})$.

Deken (1979) was able to sharpen the bounds on c_k, and also noted that as a consequence of Kingman's subadditive ergodic theorem (Kingman (1968)), that one actually has

$$\lim_{n \to \infty} L_n/n = c \qquad \text{a.s.} \tag{1.2}$$

where c depends on the distributions of the processes $\{(X_i,Y_i): 1 \leq i \leq \infty\}$.

This result naturally entails Var $L_n = o(n^2)$, but no futher progress was made on the variance problem.

The present article takes up several aspects of the study of L_n. In the second section as an elementary application of an inequality of Efron and Stein (1980), it is proved that Var $L_n = O(n)$. This makes only modest progress on the Chvátal-Sankoff conjecture that Var $L_n = o(n^{2/3})$, but it still serves to supplement (1.2) with a rate of convergence result.

The third section takes up the question of the behavior of L_n under more general assumptions than independence. A simple complement of Kingman's subadditive ergodic theorem (Kingman (1973)) is derived and then applied to L_n. The coupling method which is used here (or the Radon-Nikodym method which is sketched) may likely be of use in many other problems where subadditivity is available, but stationarity is absent.

The fourth section branches out from the explicit analysis of L_n. It addresses the question of whether there exist statistics which are more tractable than L_n, but which still reasonably measure the genetic proximity of long molecules. The principal new candidate is T_n, the total number of common subsequences. Here one can compute ET_n exactly, but we note T_n has other draw-backs to its analysis.

The final section makes brief comment on some open problems and related literature.

Acknowledgement. The observation that absolute continuity provides a second proof of Corollary 1 in Section III is due to Steve Lalley who kindly commented on an earlier draft on this article.

II. A VARIANCE BOUND

Let $S(v_1,v_2,\ldots,v_{n-1})$ denote <u>any</u> real valued function of $n-1$ vectors $v_i \varepsilon \mathbb{R}^d$; and suppose V_i, $1 \le i < \infty$, is any sequence of independent, identically distributed, random vectors in $\mathbb{R}^d$. We then define new random variables $S_i = S(V_1,V_2,\ldots,V_{i-1},V_{i+1},\ldots,V_n)$ for $1 \le i \le n$, and we further set $S_. = \frac{1}{n} \Sigma_{i=1}^n S_i$. Tukey's Jackknife estimate for the variance of S is $\Sigma_{i=1}^n (S_i-S_.)^2$, and Efron and Stein (1980) have proved the very useful inequality,

$$\text{Var}(S_.) \le E \sum_{i=1}^n (S_i-S_.)^2 \; . \tag{2.1}$$

The main point of this section is to show that (2.1) leads to the bound

$$\text{Var } L_n = O(n) \; , \tag{2.2}$$

under the general assumption that $V_i = (X_i,X_i')$ are independent, and identically distributed. In fact, one can prove the following result.

Theorem 1. For each n, suppose there is defined a function $S(x_1,x_2,\ldots,x_n)$ from $(\mathbb{R}^d)^n$ to $\mathbb{R}$. Suppose also that V_i, $1 \le i < \infty$, is any sequence of independent random vectors in $\mathbb{R}^d$, and for $1 \le i \le n$, $1 \le n < \infty$ set

$$S_{i,n} = S(V_1,V_2,\ldots,V_{i-1},V_{i+1},\ldots,V_n) \; . \tag{2.3}$$

If $E(S_{i,n} - S_{j,n})^2$ is bounded for all $1 \leq i < j \leq n$ and $1 \leq n < \infty$, then

$$(2.4) \qquad \text{Var } S(V_1,V_2,\ldots,V_n) = O(n) \ .$$

Proof. Let the bound on $E(S_{i,n} - S_{j,n})^2$ be B. Fix n, define $S_. = \frac{1}{n} \Sigma_{i=1}^n S_{i,n}$, and let

$$(2.5) \qquad D_n = S(V_1,V_2,\ldots,V_{n-1}) - S_.$$

$$= \frac{1}{n} \sum_{i=1}^{n} (S(V_1,V_2,\ldots,V_{n-1}) - S(V_1,V_2,\ldots,V_{i-1},V_{i+1},\ldots,V_n)) \ .$$

By Schwarz' inequality,

$$(2.6) \qquad \text{Var } S(V_1,V_2,\ldots,V_n) \leq \text{Var}(S_.) + \text{Var } D_n + 2(\text{Var } S_.)^{1/2}(\text{Var } D_n)^{1/2} \ ,$$

and

$$(2.7) \qquad \text{Var } D_n \leq E D_n^2 \leq B \ .$$

Since $E(S_{i,n} - S_{j,n})^2 \leq B$ one also has $E((S_{i,n} - S_.)^2) \leq B$. So inequalities (2.1), (2.6), and (2.7) entail

$$(2.8) \qquad \text{Var } S(V_1,V_2,\ldots,V_{n-1}) \leq nB + B + 2(nB)^{1/2} B^{1/2} = B(n^{1/2}+1)^2 \ .$$

This completes the proof of the Theorem with a very specific form of the $O(n)$ term ∎.

Returning to L_n we note that for $V_i = (X_i, X_i')$ and $\mathcal{G} = \{1,2,\ldots\}$ that Theorem 1 is applicable to $S(V_1,V_2,\ldots,V_n) = L_n(V_1,V_2,\ldots,V_n) \equiv L_n$. Since

$$(2.9) \qquad 0 \le L(V_1,V_2,\ldots,V_n) - L(V_1,V_2,\ldots,V_{i-1},V_{i+1},\ldots,V_n) \le 1 ,$$

it is trivial that (2.8) can be taken with $B = 1$. In summary we have the following bound.

Corollary 1. If (X_i, X_i') are i.i.d. with values in $\mathcal{G} \times \mathcal{G}$ then

$$(2.10) \qquad \mathrm{Var}\, L_{n-1} \le (n^{1/2} + 1)^2 .$$

By the usual Borel-Cantelli and subsequence arguments together with (2.9) and (2.10) one can prove a rate result.

Corollary 2. We have for all $\varepsilon > 0$ that

$$(2.11) \qquad L_n - E\, L_n = o(n^{3/4+\varepsilon}) \text{ with probability one .}$$

Since the techniques for proving (2.11) are well-known and since the result is not the best possible, there is no reason to include the proof. This is nevertheless the first rate result available on L_n, since such rates cannot be obtained in general from the subadditive ergodic theorem (c.f. Hammersley (1978), p. 670).

III. NON-STATIONARY SEQUENCES

By Deken's observation we know Kingman's theorem implies that L_n/n converges almost surely under the assumption that the V_i, $1 \le i < \infty$ form a stationary sequence. The point of this section is to give a very simple illustration of how Kingman's theorem can also be used for non-stationary processes. Naturally, one must appeal to some underlying asymptotic stationarity, but the resulting class of results seem useful enough to merit recording. In particular, one should compare the present result to the "sub-stationary" subadditive ergodic theorem of Abid (1979). That result apparently does not suffice for the application to L_n given here, and it is considerably more complicated.

By a *subadditive sequence of function* on E we denote a sequence $h_n: E^n \to \mathbb{R}$ which satisfies

$$(3.1) \qquad h_{m+n}(e_1,e_2,\dots,e_{n+m}) \le h_m(e_1,e_2,\dots,e_m) + h_n(e_{m+1},e_{n+2},\dots,e_{n+m}) \ .$$

As an example, we note that if $E = \mathcal{A} \times \mathcal{A}$ and $e_i = (a_i,a_i')$, then letting $h_n(e_1,e_2,\dots,e_n)$ denote the length of the longest common subsequence of $(a_1,a_2,\dots,a_n)$ and $(a_1',a_2',\dots,a_n')$ one has (3.1). Because of the applications we have in view, we will also focus on *monotone subadditive functions*, i.e. those functions which satisfy (3.1) as well as

$$(3.2) \qquad h_{n-m}(x_{m+1},x_{m+2},\dots,x_n) \le h_n(x_1,x_2,\dots,x_n) \text{ for all } m \le n \text{ and } \{x_1,x_2,\dots,x_n\}.$$

We will say that a stochastic process $\{X_i\}_{i=1}^{\infty}$ on the discrete state space E has a <u>stationary</u> <u>ergodic</u> <u>coupling</u> if there is a stationary ergodic process $\{X_i'\}_{i=1}^{\infty}$ on the same probability space such that $Z_i = (X_i, X_i')$ is a coupling, i.e. such that the stopping time $\tau = \min\{i: X_i = X_i'\}$ is finite with probability one.

It is well-known that couplings are a convenient and powerful way of expressing the asymptotic properties of stochastic processes (see e.g. Griffeath (1978)). The next result illustrates this ease of application.

Thereom 2. Suppose that h_n is a positive and monotone sequence of subadditive functions on E. If $\{X_i\}_{i=1}^{\infty}$ is a stochastic process with state space E for which there is stationary ergodic coupling then

$$(3.3) \qquad \lim_{h \to \infty} h_n(X_1, X_2, \ldots, X_n)/n = c \quad \text{a.s.}$$

for some constant c.

Proof. Let $\{X_i'\}_{i=1}^{\infty}$ denote the stationary ergodic process to which $\{X_i\}_{i=1}^{\infty}$ may be coupled, and let τ be the coupling time, i.e. $\tau = \min\{i: X_i = X_i'\}$. The doubly indexed process $Y_{st} = h_{t-s}(X_{s+1}', X_{s+2}', \ldots, X_t')$ is easily checked to have the properties:

$$(3.4a) \qquad Y_{su} \leq Y_{st} + Y_{tu}, \quad \text{whenever} \quad s < t < u \quad .$$

(3.4b) The joint distributions of the shifted process $\{Y_{s+1,t+1}\}$ are the same as those of the unshifted process.

and

(3.4c) The expectations $g_t = ET_{ot}$ satisfy $g_t \geq -At$ for some A and for all t.

The properties (3.4a-c) are exactly the hypotheses of Kingman's theorem (Kingman (1973)), so by its conclusion we have

$$\lim_{n \to \infty} Y_{0,m}/n = \lim_{n \to \infty} h_n(X_1',X_2',\dots,X_n')n = c \qquad \text{a.s.} \tag{3.5}$$

Here, to conclude that the limit is indeed a constant we have made use of the fact that Kingman's theorem assures that the limit is shift invariant and we have assumed that $\{X_i'\}_{i=1}^{\infty}$ is ergodic.

Now we have by (3.1), (3.2), and the definition of τ that

$$\begin{aligned} h_n(X_1,X_2,\dots,X_n) &\leq h_\tau(X_1,X_2,\dots,X_\tau) + h_{n-\tau}(X_{\tau+1}',X_{\tau+2}',\dots,X_n') \\ &\leq h_\tau(X_1,X_2,\dots,X_\tau) + h_n(X_1',X_2',\dots,X_n') . \end{aligned} \tag{3.6}$$

Since $\tau < \infty$ with probability one, (3.4) and (3.6) yield

$$\overline{\lim_{n \to \infty}}\ h_n(X_1,X_2,\dots,X_n)/n \leq c . \tag{3.7}$$

To handle the limit infimum we need only consider the analogous inequality with the variables reversed, i.e.

$$h_n(X_1',X_2',\dots,X_n') \leq h_\tau(X_1',X_2',\dots,X_\tau') + h_n(X_1,X_2,\dots,X_n) ,$$

and we obtain

$$c \leq \underline{\lim}\ h_n(X_1,X_2,\ldots,X_n)/n$$

to complete the proof ■.

Corollary 1. If V_i, $1 \leq i < \infty$, is an irreducible, aperiodic, positive recurrent Markov chain with state space $\mathcal{G} \times \mathcal{G}$ then no matter what the initial distribution $\pi(v) = P(V_1=v)$, one has with probability one

$$\lim_{n \to \infty} L_n(V_1,V_2,\ldots,V_n)/n = c$$

for some constant c.

To prove the corollary one only has to exhibit an appropriate coupling; and, in this case, the existence of such a coupling is well-known (see e.g. Hoel, Port, and Stone (1972)).

One can also prove the above corollary without recourse to coupling; one can use an absolute continuity argument. Under the hypotheses of the corollary there is a stationary measure π'. Moreover, the initial measure π is absolutely continuous with respect to π' (since by the irreducibility and positive recurrence $\pi'(a_1,a_2) > 0$, for all $(a_1,a_2) \in \mathcal{G}\times \mathcal{G}$). If $\{V_i': 1 \leq i < \infty\}$ is the process with initial distribution π' and the same transitions function as $\{V_i: 1 \leq i < \infty\}$, it is further true that the measure ρ for the infinite process $\{V_i: 1 \leq i < \infty\}$ is absolutely continuous with respect to that ρ' for $\{V_i': 1 \leq i < \infty\}$. Since $L(V_1',V_2',\ldots,V_n')$ satisfies the hypotheses of Kingman's subadditive ergodic theorem, $\{\omega: \lim_{n \to \infty} L(V_1',V_2',\ldots,V_n')/n = c\}$ is a set of ρ' measure one. By absolute continuity of $\rho << \rho'$ the set $\{\omega: \lim_{n \to \infty} L(V_1,V_2,\ldots,V_n)/n = c\}$ has ρ measure one. This is precisely the conclusion of the corollary.

IV. ALTERNATIVE STATISTICS

The random variable $L(V_1,V_2,\ldots,V_n)$ certainly is an interesting measure of genetic proximity, but it appears to be hard to handle. In such a situation it is natural to look for suitable alternatives.

To introduce one such alternative let $(X_1,X_2,\ldots,X_n)$ and $(X_1',X_2',\ldots,X_n')$ denote two sequences of values from $\mathcal{G}$. By A, B we denote subsets of $\{1,2,\ldots,n\}$, say $A = \{i_1,i_2,\ldots,i_h\}$ and $B = \{j_1,j_2,\ldots,j_k\}$ if $|A| = |B| = k$. Next we set

$$\rho(A,B) = \begin{cases}1\\0\end{cases} \text{as } X_{i_1} = X'_{j_1},\ X_{i_2} = X'_{j_2},\ldots,X_{i_k} = X'_{j_k}, \text{ or not.} \tag{4.1}$$

The statistic of interest in this section is

$$T_n = \sum_{A,B} \rho(A,B)\ , \tag{4.2}$$

where the sum is over two pairs of subsets of $\{1,2,\ldots,n\}$ and it is understood that $\rho(A,B)$ is taken to be zero if the cardinalities of A and B differ, i.e. $|A| \neq |B|$.

If the X_i, $1 \leq i < \infty$ and the X_i', $1 \leq i < \infty$ are all independent, and $P(X_i = a_j) = p_j$, $P(X_i' = a_j) = p_j'$ for all i, j, it is easy to see that

$$E\ T_n = \sum_{k=0}^{n} \binom{n}{k}^2 \left(\sum_{j=1}^{\infty} p_j p_j'\right)^k . \tag{4.3}$$

This explicit formula is quite a contrast to the mystery surrounding EL_n under similar hypotheses. A number of qualitative properties of ET_n are also evident from (4.3). In particular, if we set $p_i \equiv p_i'$ for $1 \leq i \leq |G| = a$ and take G finite, then

$$(4.4) \qquad \phi(\vec{p}) = E_p T_n = \sum_{k=0}^{n} \binom{n}{k} \left(\sum_{j=1}^{a} p_j^2 \right)^k$$

is easily checked to be a Schur-convex function, i.e. $\phi(\vec{p}) \leq \phi(\vec{p}')$ whenever $\vec{p}$ is majorized by $\vec{p}'$. (For an elaboration of this terminology see Hardy, Littlewood, and Polya (1951), and for an elaboration of the many consequences of Schur-convexity see the treatise by Olkin and Marshall (1979)).

Despite the mathematical simplicity of T_n as evidenced by (4.3) and (4.4), it provides only a partial surrogate for L_n. In the first place T_n tends to be very large, and there is no efficient algorithm for finding T_n. Thus, from a computational view point, L_n is a superior statistic. Also, as of yet, there is no information at all about the variance of T_n or of its limit properties.

V. OPEN PROBLEMS

The main open problems concern the expectations

$$\psi_n(p) = EL_n \tag{5.1}$$

under the hypotheses of independence and identical distribution as applied in (4.4).

For one explicit conjecture, it seems inevitable that $\psi_n(p)$ is Schur convex (just as $\phi(p)$ was proved to be). Perhaps it would be easier to consider the limit,

$$\psi(p) = \lim_{n \to \infty} \psi_n(p)/n . \tag{5.2}$$

Again, it must be true that $\psi(p)$ is Schur convex, but so far even this has not been proved.

The older problems concern the numerical value of $\psi(p)$. Perhaps progress can be made on this problem by taking a more algorithmic point of view. Is there an efficient algorithm for computing the approximate value of $\psi(p)$ or $\psi_n(p)$ with a guaranteed error bound?

Given the results of Section 2, it is very interesting to see if one can improve (2.10) to show $\text{Var } L_n = o(n)$. This would be the first really non-trivial step toward the Chvátal-Sankoff conjecture, and it would seem to require some genuinely new combinatorial insight to settle the point one way or the other.

Finally, the main scientific problem is to find a replacement for L_n which still has a genetic justification. The null distributions of

L_n seem like they will always be out of reach, and major progress will be made when L_n finds a suitable substitute. The statistic T_n is a reasonable first choice, but it leads to its own problems. For example, what is the order of the growth of $\mathrm{Var}\ T_n$?

In the search for surrogates for L_n, it may be critical to consider the variety of problems to which it has been applied. In addition to the application to molecule comparisons noted previously, there is a natural application in communications. In particular, Bradley and Bradley (1978) have applied L_n in the study of bird songs.

There are also a variety of potential uses in computer science and for an introduction there it seems useful to refer to the papers of Aho, Hirschberg, and Ullman (1976), Okuda, Tanaka, and Kasai (1975), Selkow (1977), and Wagner and Fischer (1974). In at least some of these papers in which L_n has been used, it seems there must exist a more tractible substitute.

References

Aho, A.V., Hirschberg, D.S., and Ullman, J.D. (1976). Bounds on the complexity of the longest common subsequences problem, J. Comput. Mach. 23(1), 1-12.

Abid, M. (1978). Un theórèm ergodique pour des processus sous-additifs et sur-stationaires. C. R. Acad. Sc. Paris, 217, Serie A, 149-152.

Bradley, D.W. and Bradley, R.A. (1978). Application of sequence comparison to the study of bird songs. Technical Report, Department of Data Processing, California State University, Long Beach, CA.

Chvātal, V. and Sankoff, D. (1975). Longest common subsequences of two random sequences, J. Appl. Prob., 12(2), 306-315.

Deken, J.G. (1979). Some limit results for longest common subsequences, Discrete Math., 26, 17-31.

Efron, B. and Stein, C. (1979). The Jackknife estimate of variance. Technical Report No. 120, Department of Statistics, Stanford University (to Appear in Annals of Statistics).

Griffaeth, D. (1978). Coupling Methods for Markov Processes, Studies in Probability and Ergodic Theory - Advances in Mathematics Supplementary Studies, Vol. 2. (ed. G.-G. Rota), Academic Press, New York.

Hammersley, J.M. (1974). Postulates for subadditive processes. Ann. Prob., 2, 652-680.

Hardy, G.H., Littlewood, J.E., and Polyá, G. (1951). Inequalities. Cambridge University Press, Cambridge.

Hoel, P.G., Port, S.C., Stone, C.J. (1972). Introduction to Stochastic Processes, Houghton Mifflin, Palo Alto, Kingman, J.F.C. (1973). Subadditive ergodic theory, *Ann. Prob.*, *1*, 883-909.

Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press, New York.

Okuda, T., Tanaka, E., and Kasai, T. (1976). A method for correction of garbled words based on the Levenshtein metric, *IEEE Trans*. Computers C-25(2), 172-177.

Sankoff, D. (1972). Matching sequences under deletion/insertion constraints. *Proc. Nat. Acad. Sci. U.S.A*. *69*, 4-6.

Sankoff, D. and Cedergren, R.J. (1973). A test for nucleotide sequence homology, *J. Mol. Biol*., *77*, 159-164.

Sankoff, D., Cedergren, R.J., and Lapalme, G. (1976). Frequency of insertion-deletion, transversion, and transition in evolution of 5S ribosomal RNA, *J. Molecular Evolution*, *7(2)*, 133-149.

Selkow, S.M. (1977). The tree to tree editing problem, *Information Processing Letters*, *6(6)*, 1-7.

Wagner, R.A. and Fischer, M.J. (1974). The string to string correction problem, *J. Assoc. Comput. Mach*., *21*, 168-173.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 293	2. GOVT ACCESSION NO. AD-A092	3. RECIPIENT'S CATALOG NUMBER 610
4. TITLE (and Subtitle) LONG COMMON SUBSEQUENCES AND THE PROXIMITY OF TWO RANDOM STRINGS		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) J. MICHAEL STEELE		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042 267
11. CONTROLLING OFFICE NAME AND ADDRESS OFFICE Of Naval Research Statistics & Probability Program Code 436 Arlington, VA 22217		12. REPORT DATE November 6, 1980
		13. NUMBER OF PAGES 16
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Common subsequence; subadditive processes; Jackknife, coupling.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

→ Let $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ and $(x'_{(1)}, x'_{(2)}, \ldots, x'_{(n)})$ be two strings from an alphabet $\mathcal{A}$, and let $L_{(n)}$ denote their longest common subsequence. The probabilistic behavior of $L_{(n)}$ is studied under various probability models for the x's and x''s strings. ←

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)